ALSO BY ROBERT WRIGHT

Why Buddhism Is True:
The Science and Philosophy of Meditation and Enlightenment

The Evolution of God

Nonzero: The Logic of Human Destiny

The Moral Animal:
Evolutionary Psychology and Everyday Life

Three Scientists and Their Gods:
Looking for Meaning in an Age of Information

THE GOD TEST

ARTIFICIAL INTELLIGENCE AND
OUR COMING COSMIC RECKONING

ROBERT WRIGHT

SIMON & SCHUSTER
New York Amsterdam/Antwerp London
Toronto Sydney/Melbourne New Delhi

Simon & Schuster
1230 Avenue of the Americas
New York, NY 10020

Portions of chapter 5, "The Elements of Understanding," previously appeared in *The NonZero Newsletter.*

First Simon & Schuster hardcover edition June 2026

SIMON & SCHUSTER and colophon are registered trademarks of Simon & Schuster, LLC

For information about special discounts for bulk purchases, please contact Simon & Schuster Special Sales at 1-866-506-1949 or business@simonandschuster.com.

The Simon & Schuster Speakers Bureau can bring authors to your live event. For more information or to book an event, contact the Simon & Schuster Speakers Bureau at 1-866-248-3049 or visit our website at www.simonspeakers.com.

Interior design by Wendy Blum

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Control Number has been applied for.

ISBN 978-1-6680-6165-7
ISBN 978-1-6682-0469-6 (Int Exp)
ISBN 978-1-6680-6167-1 (ebook)

In Memory of Frazier and Milo

CONTENTS

CHAPTER ONE

A BLAST FROM THE FUTURE

In 1983, while researching an article about artificial intelligence that I was writing for an obscure journal called *The Wilson Quarterly*, I interviewed an obscure computer scientist named Geoffrey Hinton. Hinton advocated an approach to artificial intelligence that was outside the mainstream but that, as I put it in the article, "some tout as the new wave in AI."

Four decades after my conversation with Hinton, I came across an article about him in *The New York Times* that said he was known in some circles as "The Godfather of AI." Apparently the approach he championed had worked.

In particular: This approach—after much evolution via various innovations, some of which Hinton figured prominently in—had led to dramatic progress in "generative AI," AI that creates images and sentences and other things traditionally created by human beings. In the months before the *Times* article came out, generative AI, most notably in the form of the chatbot ChatGPT, had taken the world by storm, wowing some people and freaking some people out.

Hinton, it turned out, was among the freaked out. The headline atop that *New York Times* story read "'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead." AI had over the past few years been getting smarter much faster than he had anticipated, and he was worried about the consequences.

The specific dangers cited by Hinton in the *Times* article are by now familiar. AI could (a) take lots of people's jobs; (b) make it harder to tell the real from the fake; (c) behave in unpredictably dangerous ways; and (d) be weaponized, in both literal and figurative senses, by bad actors.

And then there's (e): AI could actually *become* a bad actor—at least, a bad actor from a human point of view. It might become so smart and powerful that it could sustain itself without human assistance. At which point it might decide it had no use for human beings.

Here's how Hinton framed that prospect shortly after the *New York Times* story broke: It could turn out that "the good news is we figured out how to build beings that are immortal" and the bad news is that maybe "that immortality is not for us."

With the publication of the *Times* piece, Hinton joined the group of people known as AI doomers. And in floating that last vision of doom—the demise of our species at the hands of brainy but heartless supermachines—he joined a select subset of AI doomers that you might call "sci-fi AI doomers." They imagine AI becoming so smart and powerful as to be a kind of God—and, unfortunately, not a God that wishes us well. They imagine the human species surrendering its freedom, if not its very existence, to a ruthless superintelligence.

At the other end of the spectrum from the sci-fi doomers are people, sometimes called "accelerationists," who want AI to move along as fast as possible. They envision a world full of AI-provided wonder, a world where disease and poverty and other human afflictions are in the process of disappearing.

There's one thing that many of the more fervent accelerationists have in common with many of the more fervent sci-fi doomers (aside from fervor, I mean). And that's a belief in the coming "singularity." The singularity is the point where AI starts improving itself in a very hands-on way—amending its own algorithms to become smarter and smarter and hence better and better at amending its own algorithms. Once this "recursive self-improvement" kicks in, the rate of advance grows so rapidly that soon the term "rate of advance" doesn't do it justice; there is an "intelligence explosion," and life on Earth is forever transformed, for better or worse.

So which is it: better or worse? Who is right, the accelerationists or the doomers? The word "singularity," actually, suggests that maybe there's no way of knowing. The term comes from physics, where it is applied to black holes and denotes an impenetrably opaque boundary—a point beyond which existing theories lose their predictive value. Maybe technological change of a certain velocity could have such wildly transformative effects that there's no telling what lies ahead.

Maybe. But both accelerationists and hard-core sci-fi doomers are undaunted; when they gaze into the singularity, they think they see the future's likely contours.

WHICH AWE IS IN ORDER?

One way to get a finer feel for what they see is via the history of the word "awe."

When the word came into use, more than a millennium ago, it meant, the *Oxford English Dictionary* tells us, "immediate and active fear; terror, dread." But, the *OED* explains, "from its use in reference to the Divine Being," the word gradually came to mean "dread mingled with veneration, reverential or respectful fear; the attitude of a mind subdued to profound reverence in the presence

of supreme authority, moral greatness or sublimity, or mysterious sacredness." By the mid-eighteenth century, the word meant "the feeling of solemn and reverential wonder, tinged with latent fear, inspired by what is terribly sublime and majestic in nature." Today the word, as typically used, doesn't carry even a tinge of latent fear. To stand in awe is to stand in wonder at the greatness or power or magnitude of something, period.

Sci-fi AI doomers stand near one end of this spectrum of meaning, the end dominated by fear and dread. Accelerationists stand at the other end, contemplating the power of AI with wonder and no trace of fear. And some people stand somewhere in between—maybe cautiously optimistic, maybe somewhat pessimistic, maybe right on the border between positive and negative expectation, but in any event convinced that something momentous and transformative is unfolding.

This book has four purposes:

1. I want to convince you that if you aren't *somewhere* on the awe spectrum—if you don't feel something of great magnitude and power approaching—you aren't getting the picture. Before the artificial intelligence revolution is over, I believe, it will be the most abruptly dramatic transformation of human experience and human society in the history of our species.
2. I want to convince you that there's a way to steer the human future toward the better end of the awe spectrum—to create a world where AI's upside greatly outweighs its downside, a world where amazing things happen, a world where humans flourish—where they can find not just happiness but meaning and purpose, where they can live in freedom even amid awesomely smart and powerful machines.
3. But I also want to convince you that pulling this off—

reconciling the coming transformation with our happiness and fulfillment and freedom—will require a big reorientation of human thought. Politicians and corporate leaders and other powerful people need to reckon with the magnitude of what's coming and the specific nature of the challenge it poses. So do the rest of us, in part because so few of those powerful people seem inclined to do that reckoning.

Now here comes the weird part.

4. I want to convince you that, however cosmic the worldviews of the AI accelerationists and the sci-fi AI doomers may sound, there's a sense in which these worldviews aren't cosmic enough. In assessing the significance of the AI revolution, these people tend to put it in the context of the past century of technological change or the past several centuries of technological change. Some of them take a broader view, and locate the current moment in the whole history of technological change—which, construed broadly, encompasses more than a million years. And a few of them go even further than that. I'm with the "go even further" crowd, to put it mildly. I think we need to widen the lens by a few billion years. I think the coming of artificial intelligence marks a major threshold not just in the history of technology, and not just in the history of our species, but in the history of our planet.

Seeing AI in this context can help us envision, and shape, the future. We're definitely entering a phase of accelerating change, and it may deserve a label as dramatic as "the singularity." But one view I share with both the sci-fi doomers and the accelerationists is that this term's connotation of opaqueness is misleading. The

alternative paths that lie ahead are discernible, and the choices we need to make are clear.

Perhaps not surprisingly, given the magnitude of the coming transformation, these choices cover a lot of ground. The challenge we face is not just a political challenge—a matter of getting the laws and regulations right. And it's not just a challenge of governance more broadly—though it is definitely that. It's a challenge with a moral dimension, even what some would call a spiritual dimension.

The AI challenge will reach into our everyday lives. The decisions that each of us makes about how we use artificial intelligence, and how we think about it, are, inevitably, decisions about the role we will play in the collective human reckoning with the AI revolution. Happily, playing a constructive role in this reckoning can dovetail with using AI constructively in our own lives. Approaching the age of AI in a way that's good for you can be good for the world.

I call this whole challenge—personal, moral, spiritual, political—The God Test, for several reasons.

One of them—the sheer cosmic scale of the perspective I'm viewing the AI revolution from—will get fleshed out in chapter 3. But first, in the chapter that follows, I'll start to explore the question of why Geoffrey Hinton went from AI Evangelist to AI Cassandra. What exactly happened in artificial intelligence between 1983 and now that shocked even the person who is probably more responsible than anyone else for it happening? I'll continue that exploration in chapters 4 and 5, explaining why, for all the progress that's been made in artificial intelligence lately, we can be sure that there's much more to come.

The next four chapters, in short, are meant to convince you that you belong somewhere on the awe spectrum. Later chapters will explain how we—humankind collectively—can determine which part of the awe spectrum our species eventually winds up on.

As for the other reasons I call the coming challenge The God Test—well, these will take a while to unfold fully, but I can say this much now:

Oddly, the phase of technological evolution we're entering—a phase of ultrarapid advance that will make the futuristic feel familiar and make the familiar seem archaic—will also give new life to some ancient questions. In particular, this one: Is there some larger purpose unfolding on this planet?

In fact, this very old question is already drawing new life from the currents of technology. It is embedded in a question that has been popping up a lot in recent years: Are we living in a giant computer simulation? The question is often thrown out lightheartedly, as a kind of joke—a comment on how weird life sometimes seems. But the reason the simulation scenario gained currency in the first place is that a number of people in the tech world and its intellectual orbit, including some very smart and reflective people, started taking it seriously.

Why? Partly because of our growing power to digitally simulate complex worlds—and the recognition that there's no obvious limit to how complex such worlds could be. But I think the simulation question has also been given prominence by something subtler: a sense that our ability to simulate complex worlds, and our ability to perform other technological wonders that are coming within reach, was in the cards all along.

In other words, there's a clearer and clearer appreciation of how profoundly *directional* technological evolution is—a sense that our technologically driven history has a kind of storyline, and that we're watching the story unfold, even move toward some kind of climax. There's a sense that we're being carried on a trajectory that, whether you find it terrifying or exhilarating or something in between, is too interesting *not* to be the product of some sort of programming, some sort of authorship, some sort of primordial design.

Indeed, the whole idea of the singularity—the idea taken seriously at the two ends of the awe spectrum—carries an unspoken sense that there's a kind of narrative arrow in technological evolution, a dramatic impetus that has been pushing us toward something momentous, for better or worse.

I have no idea if we're living in a simulation, and I'm not one to earnestly invoke the word "singularity." But I do share the underlying sense that we're part of a story that's been unfolding for a long time and is now approaching a kind of climax. And I mean a *long* time: not just since the origin of human technology, but since the origin of life on Earth. And I mean a *big* climax: like, so big that, as it comes into view, it is casting our politics and our culture and our personal lives in a new light and giving them new dimensions of meaning.

The question of higher purpose is a question you can make sound new (What kind of superintelligent thing built the simulation, and what did it have in mind?) or make sound old (Is the purpose unfolding via this cosmic storyline a *divine* purpose, imbued by a God?). But however you frame the question, you wind up examining the same set of clues: Are we moving toward something better? Something worse? Is it within our power to shape that outcome? How?

That last question—the "How do we create a happy ending" question—can be put another way: What is technological evolution asking of us? What is our calling? Is it possible that the future, like gods of the past, is demanding that we move toward spiritual progress, toward a kind of enlightenment—that we become better versions of ourselves? And will we, as some of those gods mandated, pay a steep price, even an apocalyptic price, if we fail at that mission?

Among the reasons this book is called *The God Test* is that I believe the answer to those last two questions is yes. Whether

or not we live in a universe imbued with purpose, we face the kind of epic, maybe even existential, challenge that you might expect from such a universe. Understanding the logic behind this challenge—seeing the choices technological evolution is now placing before us—is, I think, the first step toward a happy ending.

CHAPTER TWO

THE GREAT INVERSION

In August of 1955, four academics—three mathematicians and a computer scientist—produced what is widely regarded as the founding document for the field of artificial intelligence. It was a proposal for a two-month workshop, to be held at Dartmouth College the following summer, that would investigate "the artificial intelligence problem." The proposal defined that then-obscure phrase like this: "For the present purpose the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving."

As for how the problem would be approached: "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."

More than half a century later, we don't know whether that conjecture is true. But we do know that, as a guide to subsequent progress in artificial intelligence, it has proved misleading. And, in the case of large language models—which power

chatbots like ChatGPT and have figured centrally in the generative AI revolution—it has proved dramatically misleading. Misleading, in a sense, by 180 degrees.

Creating large language models—LLMs—wasn't a matter of "precisely describing" the dynamics of human cognition and then replicating those dynamics in machines. The dynamics inside the machines—the inner workings of ChatGPT and all other LLM-powered AIs—turned out to be something of a mystery, even to the people who built the models.

This doesn't mean those inner workings *don't* mirror the dynamics of human cognition. In fact, findings from the field known as "interpretability"—the field devoted to figuring out what exactly is going on inside these machines—suggest that what's going on inside them does resemble, in some basic respects, what goes on inside the human mind. But most of those resemblances weren't by design. At least, they weren't by the design of human beings.

What happened, basically, is this:

Computers were given the mission of figuring out, by trial and error, how to generate coherent human language. As they got better and better at this task, useful structures of information processing developed and coalesced within the computers. And it turned out that some of those structures performed the same functions as structures within the human mind. That's why computer scientists didn't, as the Dartmouth proposal had envisioned, have to "precisely describe" the dynamics of cognition in order to build human cognitive features into machines; the machines did a lot of the building without the benefit of any such description. They in a sense reinvented some features of the human mind.

If the Dartmouth proposal had been exactly realized—if we'd built machines that think—that would have been important enough. But what we've built—machines that *invent* machines that think—is much more important.

Understanding how this happened and what it means is the key to understanding the intensity of the debate that erupted as the age of generative AI dawned. It explains why the sci-fi doomers were so terrified, and it explains why the accelerationists were so exhilarated. And it explains why both camps were at some level right: They both saw that a whole new kind of power, a once-in-a-planet's-lifetime kind of power, had been unleashed. Awe, in one sense of the word or another, was in order.

Back in 1983, when I interviewed Geoffrey Hinton, I hadn't gotten this picture. Even after I finished my research for the piece I was writing, and understood that Hinton represented a distinctive and bold vision of AI's future, I didn't understand what it would mean if his vision prevailed.

In a sense, neither did Hinton. The current power of large language models was implicit in the approach he was advocating, but only in a vague way—and only later did he, or anyone else, fully grasp that implication.

Hence the difference between the Geoffrey Hinton I interviewed in 1983—the ebullient advocate for a maverick AI research program—and the Geoffrey Hinton who has become world-famous, the somber sage who holds out some hope for the future of humankind, but says this hope can only be realized if we appreciate the very real prospect of catastrophe, even extinction. Hinton may have felt some awe back in 1983, but he didn't stand on the part of the awe spectrum he would occupy four decades later, the part where wonder is suffused with, if not outweighed by, fear and dread.

A LONG-DELAYED EPIPHANY

The piece I was working on appeared in the 1984 winter edition of *The Wilson Quarterly*. Four decades later, shortly after the release

of ChatGPT-4, I reread the piece. I also watched a lecture Hinton had given in 2018 on how modern AI models work. And for the first time, I grasped the radical difference between the kind of AI I had thought Hinton's path could lead to and the kind it had led to. I had a kind of epiphany.

The epiphany had to do with the relationship between words and meaning. The way I'd imagined computers handling that relationship had turned out to be, in a sense, off by 180 degrees.

It's not a coincidence that off "by 180 degrees" is also the way I just described the proposal for the 1956 Dartmouth conference. The mistaken expectation about words and meaning I had in 1983 is, at its core, the same mistaken expectation the authors of that proposal had.

So the good news for me is that I can honestly say a piece I wrote back in the 1980s helps illuminate a technological breakthrough that wouldn't happen until decades later. The bad news is that the illumination emanates from the piece's lack of foresight.

In the *Wilson Quarterly* piece, I outlined two rival approaches to AI. The mainstream approach was what I called the "top down" approach, which tries to give machines the power to reason by manipulating sequences of symbols—kind of like what mathematicians do when they translate a real-world problem into symbols and use a set of formal rules to work toward a solution.

The "bottom up" approach involved the building of "neural networks." Hinton and other proponents of this approach believed that true artificial intelligence would come via machines modeled on the human brain in the sense that they would have lots and lots of nodes—"neurons"—with lots of connections to one another; and the key thing wouldn't be the formal manipulation of symbols but rather (to oversimplify slightly) the patterns of interconnection that were activated for various cognitive purposes.

Because the different nodes in a neural network could do work at the same time—in "parallel"—this approach was sometimes called "massive parallelism." I still remember Hinton uttering that term energetically in his British accent—maybe because the memory would be reinforced every so often over the years when I saw his name in some article about AI. Today's large language models are very big—*very* massively parallel—neural networks.

You might ask why, given the massive parallelism of the human brain, most AI researchers back in the early 1980s weren't following Hinton's lead in their attempt to build machines that can do things the human brain does. I think part of the answer is that the *conscious* part of thinking, the part we're *aware* of, is serial, not parallel—that is, we can only think one thought at a time. But of course, these streams of linear thought wouldn't be possible if it weren't for all the parallel processing going on at an unconscious level. That's where most of the real work gets done.

The example my *Wilson Quarterly* piece gave of how a neural network might work was, I now realize, a very bad example of the kind of use Hinton had in mind for neural networks. But, for that very reason, it's an excellent example for present purposes—an example of the 180-degree misunderstanding I was laboring under back then.

MY VERY GOOD BAD EXAMPLE

The example had to do with one of many challenges our brains face when processing language: the fact that words often have more than one meaning, and which meaning is intended depends on their context. Suppose you hear someone say "John threw a ball . . ." and then pause before continuing the sentence: "John threw a ball for charity." Because of the ambiguity of the English

words "throw" and "ball," you may have gone from thinking John hurled a sphere to thinking that, actually, John hosted a dance.

In the *Wilson Quarterly* piece, I explained how a neural network might handle the challenge of ambiguity posed by that sentence. I used as my example a model that was being developed by an AI researcher named Jerome Feldman, who had earlier coauthored a paper with Hinton on neural networks.

In Feldman's model, I wrote, the two senses of the verb *to throw*—to hurl, and to host—would live in separate "nodes." Upon encountering the sentence about John throwing the ball, "both nodes would seek support for their interpretations; they would try to find other words in the sentence with which they have an affinity—with which they are connected."

To make a long story short: The "host" node would have more success than the "hurl" node, because it would have a pretty close connection to a node that represented the "dance" sense of the word "ball," which in turn would have a pretty close connection to the "charity" node. These three nodes—"host," "dance," and "charity"—would form a kind of coalition that would coalesce quickly by virtue of this closeness—and so could prevail in a kind of competition among rival sets of nodes to interpret the sentence.

If that doesn't make sense, don't worry, because how exactly Feldman's proposed neural network was supposed to work doesn't matter. Though it was an interesting model that, at a very abstract level, had something in common with large language models, the important thing for our purposes is a sense in which it was profoundly unlike those models.

Namely: With Feldman's approach, the connection between words and their meanings would have to be *built into the model by people who understood the meanings of the words*. A human being familiar with the ambiguity of "throw" would set up the "hurl" node and the "host" node and would calibrate their re-

lationships to other nodes. Even if this human somehow automated the process by feeding a dictionary into a computer, it would still be the case that the neural network couldn't start working until the preexisting human understanding of the relationship of words to meaning had been programmed into it. One way or another, the human conception of the meaning of words would have to be intentionally and explicitly implanted in the neural network.

This seemed to me, at the time, like a reasonable, even necessary approach. But it turned out that no such human tutelage would ultimately be necessary.

Today's large language models don't have any linkage between words and meanings built into them by people. When one of these neural networks, at the beginning of its training phase, starts scanning its first text, it has no clue as to what the symbols it's processing mean. And nobody ever tells it what they mean. All it can do is observe statistical patterns among the strings of symbols we call words—observe how likely one string of symbols is to follow another string of symbols given what the preceding strings of symbols were.

Yet by the end of this process—after mountains and mountains of text are scanned—the LLM can handle language as if it knew what those strings of symbols represent. Somehow, on its own, it has found a way to link words to the meanings that humans associate with them. The assumption that I and many other people were making in the early 1980s—that these links would have to be put into the machines by humans—turned out to be wrong.

In that 2018 lecture by Hinton, the one I watched in 2023, he said, "It turns out it works much better to have a system that has no linguistic knowledge whatsoever." Then he added, "That is, it's actually got lots of linguistic knowledge, but it wasn't put in by people."

THE MACHINES THAT DISCOVERED MEANING

So how did the machine manage to build up this linguistic knowledge? How did it "figure out" the meanings of words? It will be easier to answer that question after we take a look at the form this knowledge takes in the machine—the system large language models build, in the course of their training, for representing the meaning of words.

Suppose you make a graph, with an x and y axis, on which to map the words for various animals. The x axis—the horizontal axis—is labeled "degree of lethality," and the y axis is labeled "speed." Both a rattlesnake and a tiger would have pretty high x values, but the tiger would have a much higher y value. So you might plot these as: rattlesnake (8,3); tiger (10,10). And a scorpion might be, say, (5,2). You could keep adding dimensions to your map; the z axis could represent size, for example.

Large language models, during their training phase, construct a map like this, except way more complicated. For starters, the map has thousands of dimensions. In this semantic space, a word's location is specified by a very long sequence of numbers. So instead of two coordinates—like (8,3) for "rattlesnake"—you'd have thousands of numbers separated by commas (8,3,12,1,14 . . .).

Of course, it's hard to envision a semantic "space" with thousands of dimensions. Still, just as it's possible to say exactly how close "rattlesnake" is to "tiger" in two-dimensional semantic space (by measuring the distance between those two points), it's possible to say exactly how close two words are in 2,000-dimensional space. And, as you might expect, it turns out that, in an LLM's vast semantic space, words that are close together tend to have a lot of overlap in meaning. So "shoe" is closer to "boot" than either is to "tree." And "tree" is closer to "bush" than to "shoe" or "boot."

The string of numbers representing a word (8,3,12,1,14 . . .) is

called a vector, and it amounts to a way of representing the meaning of a word. If I asked you what a tiger is, you might say it's an animal with four legs and fur that has stripes and runs fast and can kill pretty big animals and eat them and reproduces sexually and nurses its offspring and so on. That list of characteristics would reflect your understanding of the meaning of the word "tiger." And each of these characteristics (including binary ones, like "has stripes") can in principle be represented along a dimension in this "semantic space" (also known as "latent space" or, more generically, as "vector space"; see "A Note on Terminology" at the end of this book for discussion of a related nuance that I'm glossing over here for the sake of simplicity).

Some psychologists think that this method of linking words to meaning—locating them in a kind of multidimensional semantic space—is the same method human brains use. But we don't really know how the brain does the linking. However, it seems safe to say *that* the brain does the linking. So this generic feature—some way of representing the meaning of words—is a feature that large language models have in common with the human brain.

That's what I mean when I say neural networks can "reinvent" features of the human brain; they build structures of information processing that perform cognitive functions performed by the brain, whether or not their way of performing that function is identical to the brain's. And this particular function—mapping words onto meaning—is, needless to say, an important one.

In a way, this seems miraculous. Think about it: You feed a machine strings of symbols that, though meaningful to humans, are, from the machine's point of view, pure gibberish. And you give the machine no information about what these strings mean to humans. And, in an important sense, the machine proceeds to "figure out" their meaning.

A FEW KIND WORDS ABOUT HUMAN BEINGS

I don't want to go overboard in my praise of these silicon brains or give short shrift to the human brains that created them. These machines didn't invent the idea of turning words into vectors, into long strings of numbers. That idea was invented by humans. Indeed, there were lots of things humans had to invent before LLMs could take shape (certainly including the "transformer," which helps the model sense important connections among words and is so valuable that it's what the *T* in GPT stands for).

Still, even if humans told the machine to use strings of numbers to represent words, they didn't tell it what those numbers should be, and they didn't tell it to use those numbers to represent *the meaning of words*. But that's what it did. The machine chose the numbers, and the numbers were meaningful.

You might even go so far as to say that machines "discovered" that meaning is a property of words—and, moreover, a property that must be recognized and represented if you're going to use human language deftly.

Computers wouldn't demonstrate such deftness until decades after my conversation with Geoffrey Hinton. But only a few years after that conversation, in the mid-1980s, Hinton published seminal papers that pointed toward the day when they would. These papers weren't about how to get a machine to generate language, and they didn't focus narrowly on the challenge of representing the meaning of words. They were about representing elements of human knowledge more broadly—representing the "features" that constitute a "concept" in various senses of that term. But in these papers Hinton emphasized two points that proved deeply applicable to language generation and helped pave the way for the large language model.

The first point was that, in a neural network, a concept could

be broken down into features, and those features could be represented in a way that was tantamount to representing them along the dimensions of a vector. The second point was that the breaking down and representing didn't have to be done by people. There were ways to get the machines to do that—to discover some of a concept's features and find a way to depict them. In a 1986 paper called "Learning Distributed Representations of Concepts," Hinton titled one section "Giving the network the freedom to choose representations." Today neural networks use this freedom to break down a concept like "tiger" into such features as "fast" and "lethal" and to represent them via dimensions in vector space.

In that paper, Hinton described an ingenious experiment that involved feeding information from family trees into a neural network and giving it tasks that would be hard to perform if it didn't manage to "discover" some features of concepts—concepts like, for example, "niece" and "nephew." One of these features was gender. Nobody had told the machine that there's something that distinguishes the people we call nieces and daughters and sisters from the people we call nephews and sons and brothers. But when Hinton observed the patterns of neuronal activity triggered by those terms, he could tell that the machine was making that distinction. There was, in effect, a gender dimension in a vector it had built—an axis on which terms like "niece" and "sister" occupied one region and terms like "nephew" and "brother" occupied another.

This neural network was, by today's standards, a primitive thing. It had ninety "neurons." You'd have to add at least four zeros to that number to approach the heft of the biggest large language models of today. But the one substantive claim I remember Hinton making during my conversation with him is that neural networks would demonstrate their value more and more compellingly as microchips got better and cheaper, and as truly massive parallel-

ism became feasible. I remember him, in that context, saying the word "massive" as if it were in italics.

And indeed, four decades later, *massively* parallel machines were wowing people. These machines were conversing in a way that seemed hard to explain unless they had some linguistic equipment that was in some sense comparable to human linguistic equipment. Yet no one had built that kind of equipment into them. Amazing.

But there's a sense in which all of this isn't so amazing, a reason we shouldn't be so surprised that these machines engineer cognitive mechanisms that do what cognitive mechanisms in the brain do. Here's the reason: The process of engineering used by the machines is at some level the same process of engineering that created the human brain in the first place. Large language models are products of evolution; they take shape by a kind of natural selection.

This is an underappreciated fact, and one reason it's underappreciated is because people commonly refer to the training of a large language model as a "learning" process. In fact, LLMs (along with image generators, facial recognition systems, and other forms of AI that use neural networks) are commonly said to be products of the "deep learning" revolution.

The word "learning" is, strictly speaking, accurate. And it's useful up to a point. It's true that an LLM, during its several months of training, learns some things that a human child learns—like the meaning of the words that constitute the particular language spoken in the child's environment. But the reason a given child can learn the local language so readily is that all children have built-in cognitive equipment, crafted by natural selection, for mastering language—such as a system for representing the meaning of words. And the LLM's version of that equipment takes shape during those same several months of training.

So, yes, the LLM is compressing large stretches of a child's learning process into a few months of training, but it's also, in a sense, compressing large stretches of human evolution into that training. Feats of engineering that took hundreds of thousands if not millions of years for natural selection to perform are performed in, relatively speaking, no time at all.

If this seems hard to envision, much less believe, please be patient. In chapter 4 I'll explain how the evolution of a large language model happens. And I'll try to give you some sense for the sweep of the evolution. Though representing the meaning of words is a vital function in the cognitive machinery we use to process language, it's far from the only function. What's more, processing language is far from the only function our brains perform to help us make sense of the world and make our way through it. And some of these other functions are now being replicated in neural networks—within the massively parallel architecture that Hinton has been advocating for more than four decades. In fact, it's not clear what functions performed by our brains *can't*, in principle, be replicated in neural networks as AI research continues—replicated by a kind of natural selection that happens in a silicon world.

One reason some people feel so much awe in the presence of AI—whether the tremulous awe of a Geoffrey Hinton or the joyous awe of accelerationists—is that they sense this. They understand how broadly applicable the power demonstrated by the evolution of a large language model is. They see the power starting to expand in range, and they see no clear limits to the expansion.

CHAPTER THREE

THE COSMIC CONTEXT

In April of 1955, a few months before the proposal for the Dartmouth artificial intelligence workshop took shape, a French priest and theologian named Pierre Teilhard de Chardin died. All of his writings, some of which had been suppressed by the Catholic Church, could now be published. This brought a burst of attention to his ideas—and not just among Christians. Teilhard had been a paleontologist and something of a visionary, and his writings encompassed evolution in the broadest sense of the term: the evolution of the universe, the evolution of life on Earth, and the evolution of technology and other products of human thought.

Figuring centrally in Teilhard's vision was a connection between technology and human intelligence, but it wasn't the same connection emphasized by the planners of the Dartmouth workshop. The connection he emphasized was connection. Teilhard was interested in how technology helped people share information and sometimes drew them into collaborative cognitive webs. He saw information technology—telephones, radios, TVs—as

linking human minds up into networks of thought that together had come to span the planet. This vision had led him to coin the term "noosphere"—based on the ancient Greek word for mind, *noos*—three decades before his death.

Teilhard described the noosphere as the "thinking envelope of the Earth" and as an emerging "planetary mind." He sometimes called it a "brain of brains"—a kind of global superbrain whose neurons were human brains.

Teilhard was writing about the noosphere long before the invention of the microchip, but he understood that computers would figure in its further development. In 1947, back when a computer with one billionth of a modern smartphone's processing power weighed 27 tons and filled a large room, he marveled at the "growth of those astonishing electronic computers which, pulsating with signals at the rate of hundreds of thousands a second, not only relieve our brains of tedious and exhausting work but, because they enhance the essential (and too little noted) factor of 'speed of thought,' are also paving the way for a revolution in the sphere of research."

But while Teilhard saw that computers would aid human thought, empowering the brains that serve as nodes in the global brain, he doesn't seem to have imagined that computers might someday rival those brains in power. He didn't anticipate the field of study that was gestating when he died. He didn't share the vision of the four people who drafted the Dartmouth proposal.

I think one key to seeing the true significance of AI—and a key to grappling with that significance successfully, guiding the evolution of AI and integrating it into our lives—is to merge the vision of those four people with the vision of Teilhard. We have to think seriously about a future in which there is something that increasingly resembles a global brain, and its neurons increasingly consist not just of human brains but of AIs. And we have to think about what kind

of brain we want that to be—about its broad contours and, in particular, about the relationship between those two kinds of neurons. There are global brains it would be great to be part of and global brains it would be miserable to be part of, and I vote for the former.

Understanding the forces that are driving us toward this hybrid planetary mind—technological forces and political forces and psychological forces—is the key to understanding two things that together are bracingly ironic:

1. Artificial intelligence makes this global brain not only more feasible but, in a sense, more necessary—necessary from the human point of view, necessary because there are no realistic alternatives to it that can safeguard our interests in the age to come.
2. At the same time, AI makes building this brain a more delicate and perilous undertaking. An AI-dependent global brain can go awry in ways the global brain Teilhard envisioned can't—and some of those ways could lead to planetary catastrophe. Others might be fine for the planet, strictly speaking, but very bad for human beings.

Before we explore this irony—explore how AI adds both momentum and danger to our construction of the global brain—it's worth pausing to understand Teilhard's conception of that momentum. It's a very vast conception. He certainly appreciated the role of technology and politics and psychology, but he also saw forces more fundamental and more foundational at work.

I don't agree with him about all these forces; I'm more conventionally scientific in my worldview than he was, especially when it comes to explaining tendencies of biological evolution. Still, I think the sheer cosmic scale of his framing of the noosphere is invaluable. If you want to understand *all* the implications of the AI